Published online 24 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D707–D711***
doi: 10.1093/nar/gkw999

New data and features for advanced data mining in Manteia

Olivier Tassy[1,2,3,*]

[1]Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Inserm U964, France, [2]CNRS (UMR 7104), Inserm U964, France and [3]Université de Strasbourg, Illkirch F-67400, France

Received September 5, 2016; Revised October 13, 2016; Editorial Decision October 14, 2016; Accepted October 17, 2016

ABSTRACT

Manteia is an integrative database available online at http://manteia.igbmc.fr which provides a large array of OMICs data related to the development of the mouse, chicken, zebrafish and human. The system is designed to use different types of data together in order to perform advanced datamining, test hypotheses or provide candidate genes involved in biological processes or responsible for human diseases. In this new version of the database, Manteia has been enhanced with new expression data originating from microarray and next generation sequencing experiments. In addition, the system includes new statistics tools to analyze lists of genes in order to compare their functions and highlight their specific features. One of the main novelties of this release is the integration of a machine learning tool called *Lookalike* that we have developed to analyze the different datasets present in the system in order to identify new disease genes. This tool identifies the key features of known disease genes to provide and rank new candidates with similar properties from the genome. It is also designed to highlight and take into account the specificities of a disease in order to increase the accuracy of its predictions.

INTRODUCTION

Manteia is a data mining system that includes several OMICs data produced for human, mouse, zebrafish and chicken. These data include functional annotations, biological pathways, protein motifs, gene expression, genetics, interactomics, molecular complexes, phenotypes and human diseases originating from various public databases (1). Data are processed upstream so they can be compared and used together across species. Manteia offers tools to explore each type of data independently but also to combine them in order to answer complex biological questions and make predictions. This can be done using a specific query language called *QueryBuilder* designed to address one or several Boolean questions to the system. This can be achieved as well by combining a mixture of independent tools using a data mining module called *Refine*. *Refine* filters the results from one tool with any other module of the system and makes it possible to get a list of genes corresponding to very specific criteria. In addition, lists of genes can be analyzed statistically to highlight the features they share using a similar approach to DAVID (2). Results can be visualized as text or using interactive graphs. Manteia is a very versatile system that can be used to analyze gene lists in many ways including the identification of genes of high biological or medical interest. *Refine* and *Query Builder* have been used in several projects to identify new disease genes using a data mining approach (1,3–5). In this new version, these tools are complemented with an entirely automated solution using a machine learning software called *Lookalike*. *Lookalike* is able to predict new disease genes based on their similarities with known causal genes in the different datasets contained in the system. This tool is also designed to analyze groups of diseases in order to highlight their specificities and use them in turn to further increase the quality of predictions. With this new release, we have also updated the gene expression module with RNA-seq and microarray data as well as the statistics module of the system with a set of tools designed to analyze lists of genes to compare their functions and to better understand their properties.

MATERIALS AND METHODS

Dataset comparison, *Batch statistics* and *Lookalike* are implemented in R 3. The web site is developed in PHP 5. The interactive graph of *Lookalike* is written in JavaScript using the D3 (data driven documents) library. Plots generated for expression data are designed using RGraph.

Expression data

In this new version of Manteia, the expression data previously based on *in situ* hybridizations (ISH) and expressed sequence tags (EST) have been replaced by RNA-seq and microarray data. These data originate from the RNA-Seq Atlas (6). They include gene expression profiles from healthy

*To whom correspondence should be addressed. Tel: +33 3 88 65 32 16; Fax: +33 3 88 65 32 01; Email: otassy@igbmc.fr

individuals for adipose tissue, colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes. The microarray dataset includes expression profiles from cancer cell lines as well. The interface developed to access these data in Manteia is based on the original tool from the RNA-Seq Atlas, offering the possibility to search for genes with a defined expression profile in different tissues and experimental conditions. In addition, this interface got enhanced in order to look for genes differentially expressed in two conditions. This is particularly useful for searching genes deregulated in tumors compared to a healthy tissue. Each query generates a set of graphs summarizing the expression of every matching gene in all the conditions and tissues available (Figure 1). This overview makes it possible to evaluate the variation of expression levels among tissues and conditions but also the consistency of these values depending on the experimental method used.

Statistics tools

The statistics module of Manteia is designed to analyze the annotation of a list of genes in order to highlight its main characteristics in the different OMICs data available in the database. This is achieved by analyzing the distribution of each annotation feature in this list compared to a reference like the genome or a given microarray (1). This analysis is performed for one annotation category at a time. However, it is often more informative to have a global picture of enriched terms for all the annotation categories available. To do this, we have developed *Batch statistics*, which is a tool that allows the user to combine annotation categories and filters to generate a flat file containing all the most representative annotations of a given set of genes in terms of functional annotations, protein motifs, phenotypes and chromosome locations.

Several results from *Batch statistics* generated from different sets of genes can be further analyzed using another tool called C*lass count*. *Class count* lists each annotation feature and reports the gene sets where it has been found. This makes it possible to identify the terms that are specific to a group of genes and list which ones are shared by one or several gene sets.

This enrichment statistics approach compares the annotation of a group of genes to a larger dataset from which it originates. In order to compare two independent datasets, we have also developed a module called *Dataset comparison*. With this new tool, the terms of a given annotation category are listed for two lists of genes. Their respective occurrences are then tested using the Fisher exact test to see if the distribution is significantly different. This is particularly useful to see if two sets of deregulated genes obtained in two different conditions are involved in different biological pathways or functions (Figure 2).

Machine learning and disease gene prioritization

In many cases, several genes can lead to the same disease or diseases that are closely related. The annotation of these genes can be analyzed to see if some rules can be learned and used in turn to detect new candidate genes. These rules can be about anything like shared biological functions, interactions and chromosomal positions. To do this, we have developed a machine learning software called *Lookalike*. *Lookalike* uses an aggregation algorithm similar to *Endeavour* (7) and *ToppGene* (8). It analyzes the annotation of known genes in several datasets from the system (Gene Ontology, phenotypes, protein motifs, chromosome distribution, sequence homology, interactome and the co-occurrence of genes in PubMed articles) to find new candidates sharing a maximum of similarities. *Lookalike* is therefore the perfect upgrade for Manteia because not only this software can utilize most of the datasets and statistics tools already developed for the system but also the aggregation algorithm is known to provide exceptional performances (9,10). *Lookalike* is very easy to use and does not require any bioinformatics or programming knowledge. The user is prompted to enter known disease genes in the training set panel of the graphical interface to run the analysis on the entire genome (Figure 3A). The search can be restricted to a list of candidates by entering their names in the candidate genes panel. More advanced options make it possible to select specific datasets to use with the algorithm, change the method used to rank candidate genes or even test the accuracy of the tool by entering known targets to see where they rank in the final classification (Figure 3A). Advantageously, *Lookalike* is integrated into a comprehensive data mining environment to evaluate and to select the best training and candidate genes possible. Because Manteia is a multi-species system, it offers the possibility to use data from orthologous genes during the prioritization. This gives the opportunity to use the mouse phenotype dataset, which is the most comprehensive of the database, to prioritize human genes, for example. The result page ranks the best candidates and shows the influence of every selected datasets in the final classification. In addition, an interactive graph allows to visually estimate the quality of a prediction and browse the results (Figure 3B). A more detailed description of *Lookalike* and its performances will be published elsewhere.

Combining statistics and data mining approaches to enhance predictions

Many specific diseases belong to a more general group like a spinal muscular atrophy belongs to the category of myopathies. The diseases from a same group can be analyzed in Manteia in order to highlight the most specific features of the targeted disease and further increase the prioritization accuracy. This is achieved by using *Batch statistics* in combination with *Class count* to identify the annotation elements that are specific to a disease compared to the other ones. Once identified, these elements are given more importance with *Lookalike* using a specific weight during the prioritization. As an example, we have computed the data corresponding to myopathy genes. We used 141 genes corresponding to six different types of myopathies (muscular dystrophies, congenital myopathies, myotonic syndromes, ion channel muscle diseases, metabolic myopathies and congenital myasthenic syndromes) according to the table of monogenic neuromuscular disorders (11). The lists of disease genes, specific features and corresponding *P*-values obtained from *Class count* can be displayed from *Lookalike*'s user interface by selecting a disease from the menu (Figure

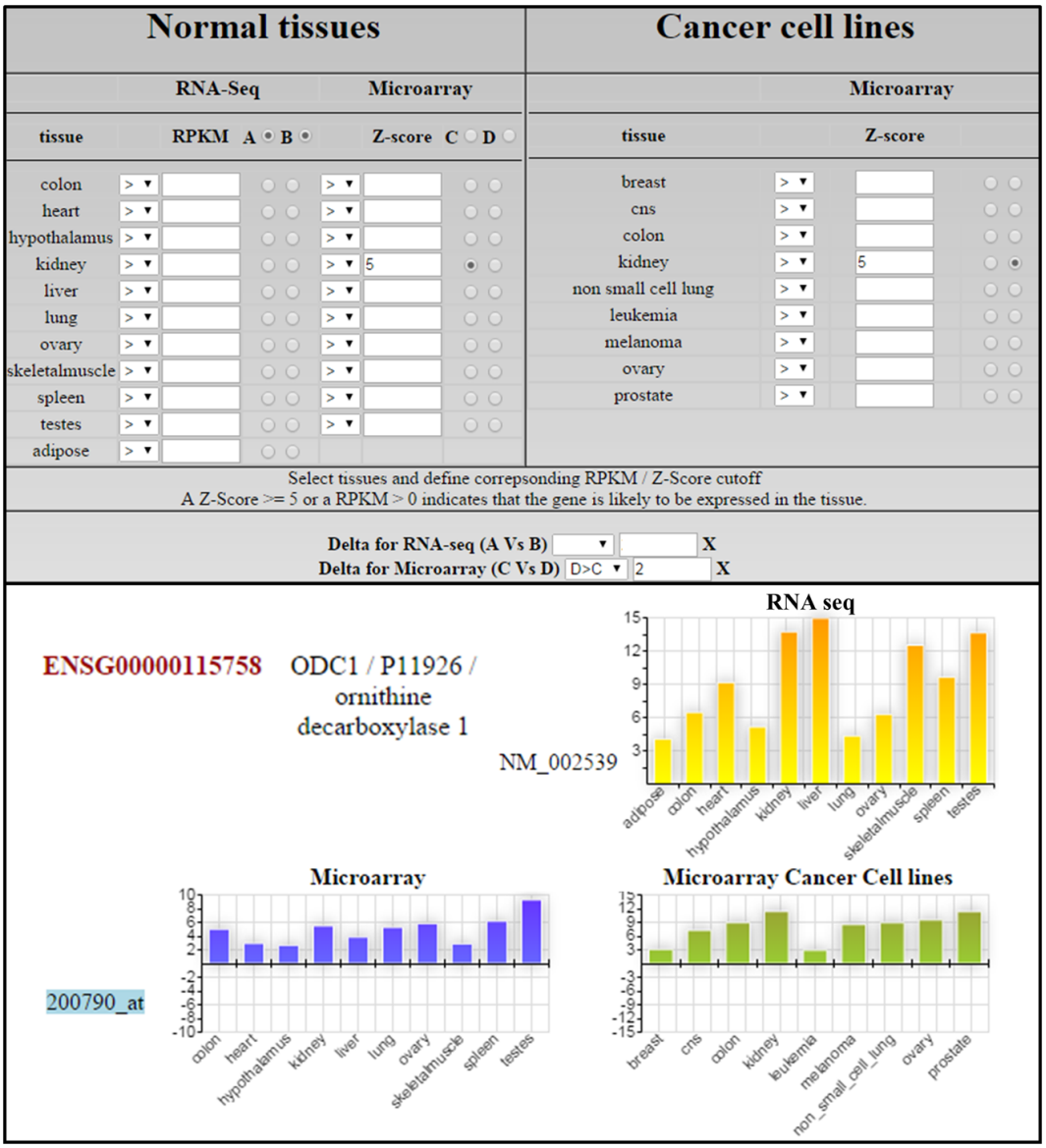
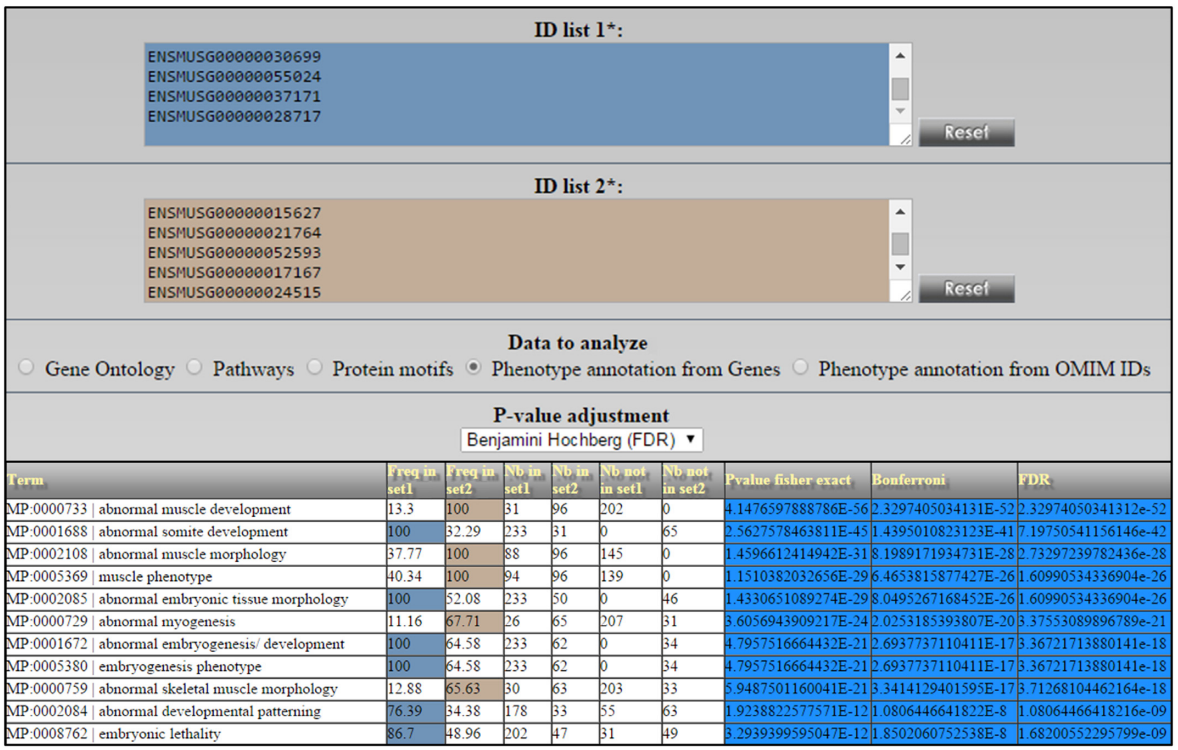


Figure 1. RNA-seq and microarray data. The RNA-seq and microarray module of Manteia makes it possible to search for genes with a defined expression profile in a mixture of tissues in both a normal and a cancer condition. In addition the radio buttons and the lower panel are designed to search for genes differentially expressed in two given conditions. The upper screenshot exemplifies the search for genes overexpressed in cancer kidney cells compared to healthy kidney cells with a minimum fold change of 2. The lower screenshot shows one of the resulting gene (*ODC1*) and the corresponding plots representing its expression levels for all the tissues in both RNA-seq and microarray datasets.

3A). This will automatically populate the different fields of the tool with the optimal settings to run the prioritization and get the best candidates related to the selected disease.

DISCUSSION

With the continuous addition of new data and tools, Manteia aims to assist scientists in their work by providing state of the art solutions to analyze their data. In this new version of the system, new microarray and RNA-seq data have been included. In addition, new ways to compare gene expression between datasets and new statistical tools are available. For many years, we have developed several approaches to prioritize genes using the information contained in the database with tools like *Refine* and *QueryBuilder*. With *Lookalike*, these tools are complemented with an entirely automated approach able to predict disease genes based on their similarities with known causal genes. However, we tried to keep this tool as open as possible so its predictions can benefit from the investigator's expertise with the possibility to manually select and weight the features of interest. Like everything else in the system, *Lookalike* can be used together with every other tools and data available. It makes it possible to constrain the candidate list to genes present on a given chromosomal region or in a given pathway, select the candidates based on a phenotypic feature instead of a disease or to further analyze the resulting candidates by accessing their annotation or using all the data mining capabilities of



Term	Freq in set1	Freq in set2	Nb in set1	Nb in set2	Nb not in set1	Nb not in set2	Pvalue fisher exact	Bonferroni	FDR
MP:0000733 | abnormal muscle development	13.3	100	31	96	202	0	4.1476597888786E-56	2.3297405034131E-52	2.32974050341312e-52
MP:0001688 | abnormal somite development	100	32.29	233	31	0	65	2.5627578463811E-45	1.4395010823123E-41	7.19750541156146e-42
MP:0002108 | abnormal muscle morphology	37.77	100	88	96	145	0	1.459661241494E-31	8.1989171934731E-28	2.73297239782436e-28
MP:0005369 | muscle phenotype	40.34	100	94	96	139	0	1.1510382032656E-29	6.4653815877427E-26	1.60990534336904e-26
MP:0002085 | abnormal embryonic tissue morphology	100	52.08	233	50	0	46	1.433065108927 4E-29	8.049526716845 2E-26	1.60990534336904e-26
MP:0000729 | abnormal myogenesis	11.16	67.71	26	65	207	31	3.6056943909217E-24	2.0253185393807E-20	3.37553089896789e-21
MP:0001672 | abnormal embryogenesis/ development	100	64.58	233	62	0	34	4.7957516664432E-21	2.6937737110411E-17	3.36721713880141e-18
MP:0005380 | embryogenesis phenotype	100	64.58	233	62	0	34	4.7957516664432E-21	2.6937737110411E-17	3.36721713880141e-18
MP:0000759 | abnormal skeletal muscle morphology	12.88	65.63	30	63	203	33	5.9487501160041E-21	3.3414129401595E-17	3.71268104462164e-18
MP:0002084 | abnormal developmental patterning	76.39	34.38	178	33	55	63	1.9238822577571E-12	1.0806446641822E-8	1.08064466418216e-09
MP:0008762 | embryonic lethality	86.7	48.96	202	47	31	49	3.2939399595047E-12	1.8502060752538E-8	1.68200552295799e-09

Figure 2. Dataset comparison. This tool takes two sets of genes as an input and lists their annotation features to test whether their distribution is significantly different in order to highlight their specificities. The color code (brown and blue) shows in which list the term is enriched. *P*-values can be corrected using Bonferroni, Benjamini Hochberg or Benjamini Yekutieli methods. Corresponding genes can be displayed for both lists using buttons on the right hand panel (not shown).

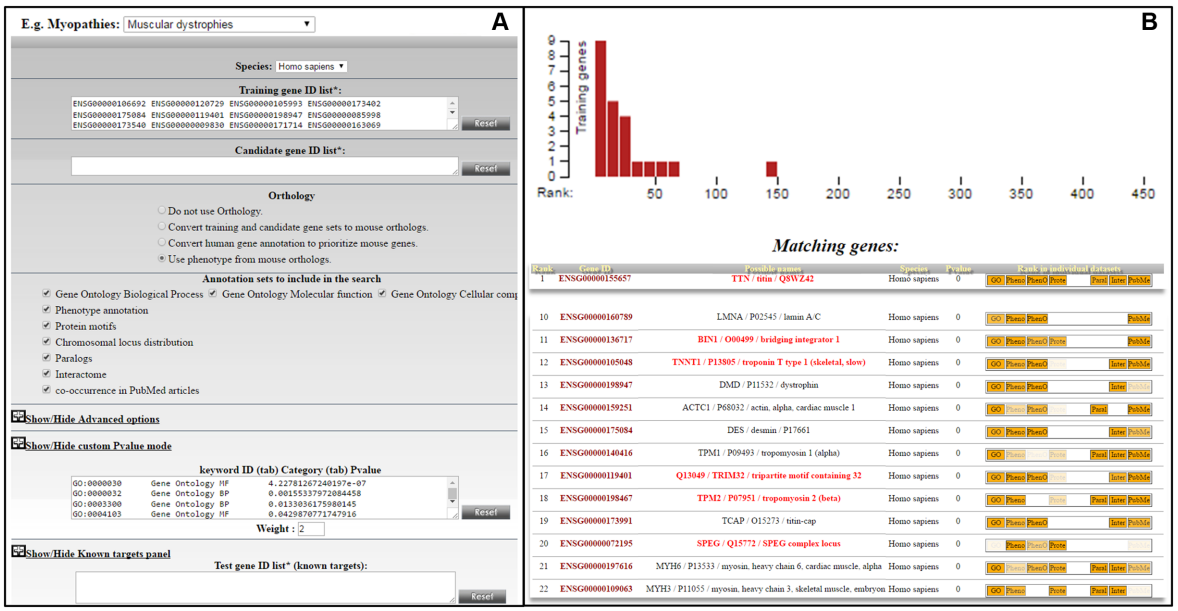


Figure 3. Disease gene prediction with *Lookalike*. (**A**) The user can use biological names or Ensembl identifiers to enter known disease genes and a set of candidates to be prioritized. If no candidates are provided, the entire genome is analyzed. By default, all the different datasets contained in Manteia are taken into account in the analysis, including the phenotype annotation from mouse orthologs. However the user can select the most relevant options using the interface. The custom *P*-value panel makes it possible to manually enter the features to be searched among the candidates during the prioritization process. It is particularly useful to focus on disease-specific features as determined by *Class count*. These fields are automatically filled with optimal parameters when selecting a myopathy from the menu on top of the page. This will perform the best prioritization possible for the selected disease. (**B**) When a prioritization is performed on the entire genome, the training genes are displayed on the graph, showing their rank distribution. These genes are highlighted in red in the following ranking list. When most training genes are found within the first positions, it means that the prioritization process picked the most relevant terms shared by a maximum of disease genes which ensure the overall quality of the prediction. The ranking of genes in individual datasets is displayed on the right hand panel with a gradient color showing their relative importance in the final prioritization.

the system. *Lookalike*'s predictions can be further enhanced by searching for the specific elements of a disease using the analyses from *Batch statistics* and *Class count*. These data are available for myopathy genes directly from the user interface and can easily be re-used to perform new candidate gene prioritizations for this type of disease. In the future, new diseases will be included in *Lookalike* to make it easier for different specialists to use this tool and discover new genes of interest. *Lookalike* and the other predictive tools of the system rely on an abundant and accurate gene annotation to provide the best results possible. The accuracy of these predictions will increase over time as more quality data will be generated by the scientific community and analyzed by the system.

ACKNOWLEDGEMENTS

I thank Olivier Pourquié and the members of his laboratory for their numerous inputs and comments about Manteia and *Lookalike*. I thank Stéphane D. Vincent, Nacho Molina, Ziad Al Tanoury and Goncalo Cadete Vilhais Neto for their careful reading of the manuscript and helpful comments and suggestions. I am also grateful to Serge Uge and Guillaume Seith for the configuration and maintenance of the servers hosting the system. Last, I thank the authors of the resources used by Manteia for sharing their work and data with the scientific community.

FUNDING

'Institut national de la santé et de la recherche médicale' (INSERM), fondation Yves Cotrel. Funding for open access charge: CERBM-GIE.
Conflict of interest statement. None declared.

REFERENCES

1. Tassy,O. and Pourquie,O. (2014) Manteia, a predictive data mining system for vertebrate genes and its applications to human genetic diseases. *Nucleic Acids Res.*, **42**, D882–D891.
2. Huang,D.W., Sherman,B.T., Tan,Q., Collins,J.R., Alvord,W.G., Roayaei,J., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, **8**, R183.
3. Baschal,E.E., Swindle,K., Justice,C.M., Baschal,R.M., Perera,A., Wethey,C.I., Poole,A., Pourquie,O., Tassy,O. and Miller,N.H. (2015) Sequencing of the gene in families with familial idiopathic scoliosis. *Spine Deform.*, **3**, 288–296.
4. Leroy,C., Landais,E., Briault,S., David,A., Tassy,O., Gruchy,N., Delobel,B., Gregoire,M.J., Leheup,B., Taine,L. *et al.* (2013) The 2q37-deletion syndrome: an update of the clinical spectrum including overweight, brachydactyly and behavioural features in 14 new patients. *Eur. J. Hum. Genet.*, **21**, 602–612.
5. Abath Neto,O., Tassy,O., Biancalana,V., Zanoteli,E., Pourquie,O. and Laporte,J. (2014) Integrative data mining highlights candidate genes for monogenic myopathies. *PLoS One*, **9**, e110888.
6. Krupp,M., Marquardt,J.U., Sahin,U., Galle,P.R., Castle,J. and Teufel,A. (2012) RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, **28**, 1184–1185.
7. Tranchevent,L.C., Ardeshirdavani,A., ElShal,S., Alcaide,D., Aerts,J., Auboeuf,D. and Moreau,Y. (2016) Candidate gene prioritization with Endeavour. *Nucleic Acids Res.*, **44**, W117–W121.
8. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.
9. Schuierer,S., Tranchevent,L.C., Dengler,U. and Moreau,Y. (2010) Large-scale benchmark of Endeavour using MetaCore maps. *Bioinformatics*, **26**, 1922–1923.
10. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.C., De Moor,B., Marynen,P., Hassan,B. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.
11. Kaplan,J.C. and Hamroun,D. (2014) The 2015 version of the gene table of monogenic neuromuscular disorders (nuclear genome). *Neuromuscul. Disord.*, **24**, 1123–1153.